1990

6. Measuring Performance in Teacher Assessment

Richard J. Stiggins
Northwest Regional Educational Laboratory

6

Measuring Performance in Teacher Assessment

Richard J. Stiggins
Northwest Regional Educational Laboratory

In performance assessment, the examinee is called upon to display behaviors or produce products which an observer evaluates in terms of prespecified performance standards. This kind of measurement plays a major role in many school evaluation contexts, including both student and teacher evaluation. The observation and evaluation of a teacher's performance in the classroom represents one of many excellent sources of information on teacher capabilities and the effectiveness of instruction. The purpose of this chapter is to review the various ways performance assessment methodology can serve in teacher-assessment contexts.

The review begins with a summary of the range of settings in which the performance of prospective and practicing teachers might be assessed, identifying the various decisions to be made and the various data sources that inform those decisions. This review of decision contexts provides a sense of the various purposes for teacher assessment and the various measurement methods used to serve those purposes. Further, it frames the role of performance assessment in a very general way.

Next, the discussion turns to a description of the basic ingredients of a performance assessment, detailing the components of

such assessments, discussing strategies for ensuring their quality, and reviewing the keys to their successful use. This description provides a sense of the range of performance assessment design alternatives available to the user.

Finally, the various purposes for teacher assessment are combined with the various performance assessment design alternatives to reveal how the two can be integrated to promote sound assessment leading to appropriate decisions in a wide range of contexts. This analysis permits an exploration of the fundamental differences in the kind of assessment methods used to achieve different purposes. In addition, it provides a means of examining strategies for ensuring valid and reliable performance assessment in each decision context.

Throughout the discussion, emphasis is placed on differentiating between sound and unsound assessment practices. The result is a detailed portrait of performance assessment at work in the arena of teacher assessment.

THE RANGE OF TEACHER ASSESSMENT CONTEXTS

There are many reasons why we assess teacher performance. Assessments inform decisions during teacher training, at the time of licensure and throughout the teacher's tenure in the classroom. In fact, we can identify at least eight specific decision contexts where assessment of teacher performance plays a role.

For instance, during preservice teacher education, instructors assess *achievement within the specific courses,* documenting the extent to which students have learned and are able to apply the pedagogical principles covered. Assuming that (a) the content of each course contributes to the development of a competent teacher and (b) course assessments cover the content taught, each course assessment verifies achievement of essential pedagogical knowledge and/or skills. Also during preservice training, faculty evaluate the performance of students as they participate in field experiences, such as their *student teaching experience.* Whether evaluating course achievement or student-teaching performance, the use of sound assessment methodology is very important. In either case, sound assessments are those that sample teacher performance in a representative manner according to clearly specified course outcome and/or classroom performance criteria.

Teacher certification and licensing, another teacher-assessment

context, calls for the new teacher to be licensed to practice, often on the basis of course work completed and examination. Such evaluations often are required by state law and are carried out by state licensing boards. States often require that the teacher who has been given an *initial certification* be reevaluated at a later time to receive a *permanent certificate.* Because these are important decisions, it is incumbent upon the assessor to employ the soundest of assessment methods. Those methods should evaluate the important dimensions of good teaching based on (a) completion of an appropriate professional preparation program and (b) observation and evaluation of a representative sample of the teacher's classroom work.

Local school districts also assess teachers for a variety of purposes. For instance, they assess prior performance in some terms for purposes of initial *hiring.* That is, the candidates for particular teaching positions are ranked for selection on the basis of some criteria. Further, districts evaluate teachers periodically to be sure *minimum teaching competencies* are being demonstrated in the classroom. Such evaluations often are required by state law and/or collective bargaining agreements. This assures the public that only competent teachers teach. Because important decisions rest in the balance, both the teacher and the district count on the use of the best available assessment methods to assure appropriate decisions.

Yet another context in which districts assess teacher performance is to promote the *professional development* of teachers. For example, assessment for development is becoming more and more common during the induction of new teachers, when many work with mentors to learn the ropes. Mentors assess performance and provide feedback. Assessment for professional development also comes into play when practicing teachers are judged not to be measuring up to minimums. They are asked to participate in important professional development activities. These activities are accompanied by careful repeat evaluations to be sure the needed changes come about. In addition, a great many teachers who may be outstanding professionals already are keenly interested in pursuing ongoing professional development. As these teachers pursue their growth-producing goals and activities, their professional development can be monitored to document and facilitate improvement.

And finally, it has become more and more common for districts to assess and make decisions about teachers' *career advancement* to higher levels of a career ladder, such as to mentor teacher status, or higher pay levels. In this case, the objective of assessment is to

TABLE 6.1
Relationship Between Teacher Assessment Contexts
and Associated Measurement Methods

	Measurement Methods					
	Paper & Pencil Test					
Decision Context	*Basic Academic Skills*	*Subject Matter*	*Pedagogical Knowledge*	*Course Completion*	*Interview*	*Performance Assessment*
1. Preservice teacher education						
A. Measuring course achievement	N	T	T	N	P	P
B. Evaluation of student teaching	N	N	N	N	P	T
2. Teacher Licensing						
A. Initial certificate	T	T	T	T	P	P
B. Permanent certificate	P[a]	P[a]	P[a]	T	P	T
3. District evaluations						
A. Selection to be hired	P[b]	P[b]	P[b]	T	T	P
B. Assurance of minimum competence	P	P	P	N	T	T
C. Assessment for professional development	N	P	P	P[c]	P	T
D. Career advancement	N	P	P	P	T	T

Note. T = Typical in this context; P = Possible but rarely used; N = Not appropriate

[a]More appropriately done earlier

[b]If test has been validated to assure job relevance

[c]If courses are intended to contribute to specific professional development goals

focus the assessment far above minimum competence and beyond individual professional development to the identification of those who have attained the highest levels of professional excellence. Because few typically are offered the opportunity to ascend to these levels, it is incumbent upon the assessor to screen candidates very carefully using the highest quality data in teacher performance.

Thus, there are a great many contexts for teacher assessment and a great many decisions to be made on the basis of the results. Some of the assessments can be and typically are based on paper-and-pencil tests of basic skills, subject-matter knowledge, and/or pedagogical knowledge. Others often are based on a review of training program transcripts and records. Still others rely on interviews. But many, in fact most, of these decision contexts rely at least in part on performance assessments; that is, measurement of teacher performance based on observation and evaluation of teachers' classroom behaviors and associated products. The relationship between the various decisions and alternative measurement methods used is outlined in Table 6.1. In five of the eight contexts, performance assessment is the typical measurement method used. In the other three, performance assessment has a potential role to play. For this reason, a basic understanding of performance assessment methodology is essential for those who would assess teachers.

PERFORMANCE ASSESSMENT OVERVIEW

To understand the role performance assessment can play in these teacher assessment contexts, we need to begin with an understanding of the basic ingredients in a performance assessment. Any such assessment can be described in terms of four key components: the purpose(s) for assessment, performance to be evaluated, exercise(s) that elicit performance, and performance rating procedures.[1] Each component contains within it a number of subcomponents, or design alternatives, from which the user can choose. It is through an examination of these design options that we can fully understand the wide range of performance assessment possibilities.

---

[1]For a more detailed treatment of these ingredients, see Stiggins, R. J. (1987). Design and developing performance assessments. *Educational Measurement: Issues and Practice, 6*(3), 33–42.

The Basic Components. As this section unfolds, it will become clear that performance assessments vary greatly in their form. The form of any particular assessment depends on its purpose. Without knowledge of the purpose, it is virtually impossible to design a useful assessment. Consequently, the first step in the design of a performance assessment is *to specify the purpose.* The range of alternative purposes is quite wide. For example, performance assessments can inform decisions about examinee strengths and weaknesses, rank order examinees for selection, certify mastery of minimum competencies, and/or evaluate the impact of some instructional treatment of a group of students. Some of these decisions require the generation and use of criterion-referenced data, whereas others rely on norm-referenced results. The type of data needed, in turn, will influence the type of performance-rating procedures used. Thus, the design of a performance assessment begins with the specification of purpose: What is the decision to be made and who is the decision maker?

Step two in the design of a performance assessment calls for the *description of the performance to be evaluated.* Two aspects of this are important. First, the designer must specify the type of performance to be observed. Three options are available. One can observe examinee behaviors, products created by the examinee, or some combination of these two. The option selected in any particular case is a function of where the best, most conveniently accessible evidence of proficiency can be found.

Second, the designer must specify the performance criteria or the dimensions of performance to be rated—in observable terms. Each key dimension must be specified in two parts: a definition and a performance continuum. Without a clear vision of performance, including a sense of the difference between poor and outstanding performance, it is impossible to judge proficiency. Because these criteria can form the basis of a number of important decisions, no single performance assessment design factor deserves more careful thought than this one. Clearly articulated criteria are an absolute necessity.

Step three in the development of a performance assessment calls for the *design of assessment exercises.* This step includes three design decisions. First, the user must specify the form of the exercises. Three options are available: structured exercises, natural events, or a combination of the two. Structured exercises are specific tasks or problems presented to the examinee to be completed or solved. They are fabricated to simulate part of the actual performance arena and are most useful when (a) the decision context

requires that all examinees have an equal and standard opportunity to demonstrate their proficiency or, (b) a natural context for observation simply is unavailable or impractical. Observation of naturally occurring events can serve as the basis of performance assessment when those events contain sufficient observable evidence of proficiency to allow for valid and reliable evaluation. Such naturalistic observations often provide the most valid assessment of performance because they often provide a high-fidelity representation of the real world. However, natural events also are less subject to tight control and standardization. The requirements of performance in natural contexts can vary from occasion to occasion.

Step three in the performance assessment design process also requires the determination of the examinee's level of awareness that an assessment is taking place. A performance assessment can be open and public, unobtrusive or some combination of the two. The option of choice depends on the purpose for the assessment. Unobtrusive assessment, although not common, can be a useful tool. It often requires the observation of naturally occurring events and can be an advantage when (a) the examiner is interested in evaluating typical—not best possible—performance, and/or (b) the examinee is troubled by debilitating evaluation anxiety. Unobtrusive assessment can help alleviate the anxiety.

Finally, step three requires the specification of the number of samples of performance to be observed before making a judgment regarding the adequacy of performance. Choices include one sample of performance at one time, multiple samples at one time, and multiple samples over a period of time. In this case, the option of choice is a function of the seriousness of the decision to be made and the amount of time required to make one observation. The more serious the decision, the more evidence is needed. The more time required per observation, the more expensive will be the assessment. Thus, the context determines the amount of data gathered. In many cases, the choice of options turns on which alternative covers the range of situations in which the examinee might be called upon to demonstrate proficiency in the future.

The fourth and final step in the design of a performance assessment is the development of *performance rating procedures*. The user must determine the nature of the score(s) needed, the identity of the rater(s) and the recording method. With respect to the score, one can opt for a holistic rating (one score covering overall performance), an analytical rating (each performance criterion rated separately), or some combination of the two. This choice is a func-

tion of the nature of the data needed to make the decision at hand. For instance, selection decisions often require that examinees be ranked on the basis of overall performance, whereas diagnostic decisions require more detailed analyses of performance. For the former, a holistic score may suffice. For the latter, analytical data are needed.

When selecting the rater(s), the user has four options: independent rating by expert judge(s), peer-rating, self-rating, or some combination of these. Independent expert judges are needed when ratings require specialized knowledge or expertise, a competitive decision requires equal opportunities for all, and/or examinees have a vested interest in results and may be perceived as benefiting unfairly from self- or peer-rating. On the other hand, peer- or self-ratings are viable options when examinees are capable of learning and applying the performance criteria, have nothing to gain from inflating or deflating their ratings, and/or have time to observe and rate performance.

Finally, the recording method can take various forms, such as a checklist (list of attributes to be marked as present or absent in performance), a rating scale (continuum on which poor to outstanding performance is rated), an anecdotal record (verbal description of performance), a portfolio (file of samples of products), an audio or videotape or some combination of these. Another often used, but completely unreliable and unacceptable option is mental record keeping.

Each of the acceptable options has advantages and limitations. Checklists and rating scales provide an efficient means of evaluating and recording but can result in somewhat superficial pictures of performance. Anecdotal records and portfolios provide more detail but are cumbersome. Tapes provide a detailed record of performance, but contain no ratings. The best choice is a function of the use to be made of the results.

In this section, we have reviewed several specific design decisions to be made by the developer of a performance assessment. Each contained a number of design alternatives within it, as summarized below:

1. Specify the assessment purpose
   A. Identify decision(s) to be made
   B. Identify decision makers
2. Define performance to be evaluated
   A. Select the type of performance
      1) Behavior

2) Product
3) Combination of behavior and product

B. Specify performance criteria
1) Define each
2) Develop the performance continuum for each

3. Design exercises
A. Select the form
1) Structured exercises
2) Natural events
3) Combination of structured exercises and natural events

B. Determine examinee's level of awareness
1) Public
2) Unobtrusive
3) Combination of public and unobtrusive

C. Determine number of samples to be observed
1) One sample, one time
2) Multiple samples, one time
3) Multiple samples over time

4. Develop the performance-rating plan
A. Select the type of score
1) Holistic
2) Analytical
3) Combination of holistic and analytical

B. Identify the rater
1) Expert
2) Peer
3) Self
4) Combination of above

C. Select the recording method
1) Checklist
2) Rating scale
3) Anecdotal record
4) Portfolio
5) Audio or videotape
6) Combination of above

Because these ingredients can be assembled in so many combinations, the range of possible forms of performance assessment is very broad. This is precisely why this kind of assessment is so valuable. In the next section, we explore how it can be molded to fit a wide range of teacher-assessment needs.

Ensuring Quality. Before we go on to explore the specific application of this methodology in various teacher assessment contexts, however, we need to review the keys to successful assessment in general. The quality of our measurement of performance is assured by giving careful attention to the purpose for assessment, communicating effectively, attending to the validity of the assessment and maximizing the reliability of the assessment.

Careful attention to purpose provides a clear sense of the decision to be made and the decision maker. With these factors in mind, the user can tailor exercises, performance criteria and performance records to fit the context. In the absence of a clear sense of purpose, it is impossible to design a meaningful or useful assessment.

Clear communication is central to effective assessment in the sense that examinees must understand the performance expectations or criteria and they must understand the exercises to which they must respond and the feedback provided. Of course, this requires that the assessment developer pay careful attention to these factors during the design process. Clear and thoughtful communication is also central to the effective delivery of feedback to examinees regarding their performance. The most appropriately designed assessment can become completely ineffective if results are poorly communicated.

In order to maximize the validity of the assessment, we must thoroughly articulate the performance dimensions to be evaluated and appropriately sample the range of possible performance arenas. Performance dimensions must be clearly defined to be effectively evaluated, and each dimension must be accompanied by the specification of different levels of performance or a continuum reflecting poor to outstanding performance. These specifications, along with a set of exercises that fairly and thoroughly samples the range of instances when the examinees might be called upon to use their skills, contribute to a valid assessment.

Finally, the keys to a reliable assessment are clear criteria used by thoroughly trained raters in the context of a carefully articulated scoring process. The acid test of the objectivity of performance ratings is for two independent judges to assign the same rating (within a small margin) to the same sample of performance. If such ratings vary greatly, then unclear criteria, rater bias, and/or inadequate rater training are indicated. Also central to reliable assessment are uniform assessment conditions and a large enough sample of performance to lead to confident judgment. The more serious the decision, the more data must be gathered.

Thus, in general, sound performance assessments are those that have a clear purpose, rely on clear communication, and are valid and reliable. Adherence to these quality-control standards can render these assessments more useful in many contexts than any other type of assessment. Let's explore why this is particularly true in teacher assessment.

MEASURING PERFORMANCE IN TEACHER ASSESSMENT

In the introduction, eight specific contexts were identified in which prospective or practicing teachers are assessed for purposes of making specific career-related decisions (see Table 6.1 for review). Further, we pointed out that any of a variety of assessment methods might be used to gather needed information. In five of the eight cases, performance assessment is the typical mode of measurement used. In the remaining three performance assessment represents a viable option. Having defined the active ingredients in a performance assessment and reviewed the wide range of design alternatives, let's explore how those design choices can vary with the teacher assessment context.

To reach this goal, the basic performance-assessment structure (purpose, performance, exercises, and rating procedures) is used to profile assessment practices in each context (course evaluation, student teacher evaluation, initial certification, permanent certification, hiring, minimum competency evaluation, professional development and career advancement). This analysis and profiling process allows us to highlight the keys to conducting valid and reliable assessments in each context.

Measuring Course Achievement. As prospective teachers complete their preservice training and as practicing teachers complete graduate course work, their achievement of course objectives is measured to allow instructors to diagnose student needs and/or determine course grades. Because completion of specific required courses is often a matter that bears directly on later career-related decisions for teachers, these course assessments are very important. Although many such assessments are based on paper-and-pencil tests and performance assessments are rare, they remain an excellent option.

In defining the performance to be evaluated, the instructor needs to consider the intended outcomes of the course. Those out-

comes might be reflected in the trainee's ability to demonstrate a particular instructional method in a simulated classroom setting (an observable behavior) or in the trainee's ability to create a specific product, such as a lesson plan or a test developed to reflect a particular unit of instruction. The performance criteria need to be defined in terms consistent with intended course outcomes and in sufficient detail to reflect the specific behavior(s) or product attribute(s) to be demonstrated. Such detailed criteria are central to the teaching of instructional methods. *Without such criteria clearly in mind, it is impossible to teach, let alone measure, mastery of important instructional skills.*

As the instructor develops performance exercises to evaluate college-course achievement, he or she would probably rely on simulations and other college classroom activities to provide needed evidence of proficiency. Assessments in this context probably are announced and may provide only a few samples of the performance of individual students if a large number of students is enrolled in education courses. Naturally occurring classroom events are less likely to be available for the course instructor to use as a basis for evaluating student achievement. However, the more opportunities undergraduates have to demonstrate new skills in real school classrooms, the better will be the diagnostic quality of the performance ratings that may come from classroom performance assessments, and the more effective will be the quality of teacher training. Furthermore, the more opportunities instructors have to watch their students use the skills learned in courses, the better will be the teacher training programs.

Rating procedures used in course assessments will vary as a function of the specific decision to be made. If the reason for assessment is to diagnose the needs of students, then analytic data are required. However, if the reason is to assign a grade, an overall rating of performance may suffice. In most cases, the rater of performance will be the instructor. However, peer- and/or self-rating also remain options that offer the distinct advantage of allowing students to learn and apply the performance criteria to their own and each others performance, which in turn often enhances the performance of the student rater. The choice of record-keeping systems to use is a function of the nature of the depth of information needed. General grading decisions might be based on checklists and ratings, whereas diagnostic decisions may require a detailed analysis of anecdotal events, a portfolio of products or videos of students in action.

The keys to gathering a valid and reliable data on of teacher

performance in the context of assessment during teacher training are to:

- be sure to use performance assessment when it represents the best way to measure intended course outcomes,
- translate intended course outcomes into clear and detailed performance criteria, and
- carefully train raters to apply the criteria,
- gather enough observational data to provide a representative picture of the performer's proficiency in all relevant classroom situations.

In other words, validity and reliability problems arise when we:

- use paper-and-pencil tests to measure traits better reflected in student behavior and/or a student product,
- teach and assess on the basis of ill-defined criteria, and
- fail to sample the full range of relevant student capabilities.

Trained raters using clear, course-relevant criteria to rate an appropriate sample of performance can conduct high-quality assessments in teacher training programs.

Evaluating Student Teaching. Within the context of teacher training perhaps the single most important teacher assessment is the evaluation of the students performance during student teaching. In this case, the typical mode of measurement is classroom observation and judgment. Although paper-and-pencil tests, interviews, and the like may play a role, no mode of assessment rivals performance assessment as the means of evaluating success in this context. The decision to be made is a very important one: Is this teacher ready to take responsibility for a classroom? This is the capstone evaluation that reveals whether the student is able to assemble all of the ingredients in sound instruction. But we must also keep in mind the fact that student teaching represents an extension of the preservice teacher-training experience. Trainers have an obligation to observe, diagnose and help improve the practice of those whom they oversee. The decision makers in this context are the supervising teacher and the college coordinator of student teaching.

In defining the performance to be evaluated, the evaluators often rely on an examination of behaviors and products. But in this

case, the performance criteria are far more complex than those covered in any single course. In this case, the criteria need to cover all of the important competencies needed to be able to take responsibility for a classroom. They need to reflect dimensions of performance in instructional design (e.g., ability to use technology effectively), classroom management (able to control time on task), assessment and evaluation (apply a range of testing methods), subject-matter knowledge, and other key areas. In a sense, they may represent a compilation of the various course outcomes. In addition, and this is crucial, they may not be the same for all student teachers. Rather, the attributes of a good teacher and good teaching may vary by grade level, subject matter, and school context. If they do, performance assessments must be sensitive to those differences. But whatever the dimensions of good teaching, those dimensions must be thoroughly defined and a continuum of performance must be articulated for each.

In the student-teaching context, the evaluation is based on observation of naturally-occurring events with the student aware that an assessment is underway. Because the field experience is many weeks long, multiple samples of performance are gathered over time. For these reasons, student teaching represents an excellent opportunity for students to learn via detailed feedback from supervisor teachers and to demonstrate their mastery of the skills of teaching.

If the evaluation of performance during student teaching is to fulfill its potential, clear and appropriate performance criteria must be applied by trained supervisors to provide diagnostic (analytical) information leading to a sound overall (holistic) judgment of proficiency. Throughout the field experience, evaluators might use checklists and ratings scales tailored to provide unique information to the student regarding how to improve. In addition, anecdotal records and portfolios might be kept by the student and the supervising teacher as evidence of skill and accomplishment. Audio and videotapes also can be valuable tools in this context.

Thus during student teaching, performance assessments of many types might come into play. But to be valid and reliable, once again, they must be:

- based on clear criteria developed to reflect the full range of appropriate teaching skills
- evaluated through the observation of the full range of classroom events and examination of all relevant documents and artifacts

- by multiple trained observers and raters
- who spend enough time observing and providing feedback to promote the development and appropriate certification of the new teacher.

In the student-teaching context, problems of dependability arise when criteria are vague and/or inappropriate, observations fail to sample enough classroom performance to produce generalizable results, raters are untrained or careless, and/or student teachers receive information that fails to provide guidance as to how to improve.

Initial and Permanent Certification. Upon completion of an accredited program of study, the new teacher is eligible for initial certification to practice. Most often, this is a temporary certificate allowing the teacher to practice for some specified period of time. At the end of that period, if additional training experiences have been completed, a permanent certificate is granted. In this context, then, the first decision is whether to grant a license to teach, and the follow-up decision is whether to allow the teacher to retain that license.

In most cases, these decisions are made by a state bureau based on (a) the analysis of transcripts of courses completed by the teacher and (b) performance on basic academic skills tests. Thus, the assumption is made that the courses taken actually teach required teaching skills and the assessments within those courses test the students' mastery of important skills. If the courses have been carefully developed on the basis of a task analysis of good teaching, the training programs have included a sound student teaching experience, and assessments have been designed to test both knowledge and performance, this assumption may be valid. However, if the basic principles of sound performance assessment have not been adhered to in the evaluation of course and student teaching achievement, transcript analysis will not contribute to valid licensure decisions.

To the extent that certifying agencies are uncertain regarding the quality of undergraduate- and/or graduate-course assessments of teaching skills, supplemental performance assessments might be added to data informing the initial certification decision. These might take the form of precertification internships in which new teachers perform under the watchful eye of an experienced mentor, the successful completion of an employment experience (see minimum competency assessment following), or the successful

completion of assessment-center simulations of particular classroom events. In any case, the rules of evidence for sound performance assessment must be adhered to: clear and appropriate performance criteria, a representative set of exercises, and carefully planned and conducted evaluations of performance.

Hiring. When school districts have positions to fill, they must screen candidates, rank order them, make sure the top ranked satisfy minimum acceptance standards, and select the most qualified to be hired. In most cases, the selection process is the responsibility of the district personnel director and the building principal in which the new teacher will be employed. Further, most such decisions are based on the review of placement papers (including transcripts and recommendations) and personal interviews with the teachers. Such decisions rarely are based on the demonstration of teaching skills by the prospective employee. Reliance on records, recommendations, and interviews as information sources for hiring rests on the assumptions that:

- course and student-teaching assessments dependably tested appropriate classroom performance,
- those providing recommendations have observed and evaluated classroom performance in an appropriate manner,
- prior supervisors of experienced teachers have conducted sound performance assessments, and
- the self-reports of performance capabilities presented by candidates during interviews accurately reflect true capabilities.

If these conditions are met, then an independent demonstration of teaching skills prior to hiring is unnecessary.

In short, it becomes obvious at the time of hiring that teacher assessments and decisions build upon one another. Each assumes (often without verification) that the preceding performance assessments were sound. Is this a defensible assumption? In many cases, it may not be. The quality of course and student-teaching assessments is generally unknown. Often, those assessments are designed and conducted by instructors largely untrained in performance assessment methodology. Further, the quality of the performance observations and evaluative judgments of principals and other supervisors has been called into question in the research

literature.[2] We know that many of these assessments are conducted by staff untrained in proper methods of observation and evaluation. Finally, we know that self-assessments, although very useful in some contexts, probably are biased in an employee selection situation. Clearly, there is reason to question at least some of the performance assessment methodology that is assumed to support the hiring decision.

Therefore, it may be useful to consider other performance assessment options. The simplest alternative might be to sprinkle interviews with descriptions of typical classroom problems designed to probe the prospective teacher's plan of action in dealing with hypothetical instructional design and classroom management problems. A second option might be to present more complex simulations via video or role play for the candidate to address through the development plans of action. Yet a third choice might be for the teacher to develop products, such as lesson plans or paper-and-pencil tests, which might be evaluated in terms of specific attributes. Another possibility is for the teacher to take over a classroom as a substitute for a day to demonstrate instructional skills. Finally, teacher applicants might prepare videos demonstrating their competence for presentation to personnel officers as part of their application for employment.

In all cases, it will be the responsibility of the assessor to develop a clear and appropriate set of performance criteria, create a range of exercises to sample relevant job-performance situations, train multiple raters to apply criteria appropriately, and conduct the performance assessments in a fair and consistent manner so as to assure all candidates an equal opportunity to be selected. The selection context involves very high-stakes decisions for the teachers involved and for the students in the classes of the teacher hired. For this reason, it is essential that the decision be based on dependably evaluated demonstrations of classroom skill. If records, recommendations and interviews cannot be counted on to provide the needed performance information, alternative performance assessments should be considered.

Assessment of Minimum Competence. Once teachers are hired and begin to practice in the classroom, the assessment of their

[2]This literature is summarized in Stiggins, R. J., & Duke, D. L. (1988). *The Case for a commitment to teacher growth: Research on teacher evaluation.* Albany, NY: SUNY Press.

performance via observation and judgment begins in earnest. Supervisors often are required to carry out regularly scheduled evaluation procedures spelled out in state law and local collective bargaining agreements. Their immediate purpose is to make personnel management decisions: Who will be released? Who will be retained? And, ultimately, who will be granted tenure in the district? The immediate goal of these evaluations is to manage incompetence; that is, to see that only competent teachers remain in classrooms. Incompetents must be trained or removed. Such evaluations recur every year or two in most districts. For the teacher who continues to demonstrate minimum competence, these evaluations will not have real impact on the teacher. However, for the teacher who is having difficulty, the evaluation will result in a call for improvement and, if that fails, dismissal.

Because dismissals can be contested, performance assessments used in this context must be legally defensible. Evaluation must be conducted so as to protect the due process rights of the teacher and the district. Therefore, the performance criteria must reflect the competencies that comprise minimally competent teaching. Further, the criteria must be standardized for all to assure equal opportunity for all to succeed. Finally, they must be public and available for all to see and understand in advance of any evaluation.

Sound practice suggests that the performance assessment be based upon observation and evaluation of both classroom practices of the teacher and documents created by the teacher in support of the instructional effort (e.g., lesson plans). Thus, the assessment is based almost completely on the observation of naturally occurring events and products. If the supervisor adheres to the letter of the law and the strict wording of the collective bargaining agreement, the performance assessment often is based on one or two 1-hour observations during the school year, scheduled at mutually agreeable times. Although specific requirements may vary from district to district, observations and evaluations often are made by one person (the supervisor) and ratings reflecting each of the minimum competencies are combined into an overall rating, which is communicated to the teacher in writing. If the judgment is that a particular competency has not been met, a plan of action is developed to overcome the deficiency. If no deficiencies are noted, no action results.

The keys to valid and reliable assessment in the context of minimum competency evaluation are:

- reliance on clearly stated, job-relevant minimum competencies,
- evaluation of performance in terms of a broad range of naturally occurring classroom events and artifacts
- evaluation of performance by observers trained to apply the performance criteria in a systematic and consistent manner,
- the delivery of feedback on performance with an opportunity to overcome any deficiencies, and
- the maintenance of records detailing the teacher's performance rating on all relevant competencies.

Difficulties arise when the performance expectations are not specified at all, are stated in broad, vague terms, or fail to reflect the specific task demands placed upon the teacher. Further, the dependability of assessments based on one or two prearranged samples of teacher performance over the span of a year or two or three must be questioned. The extrapolation from such scant and possibly biased data to the thousands of hours of teaching conducted by the teacher is indefensible. In addition, the danger of bias is great when an evaluation of teacher performance is based on the observations and judgments of only one person. These problems are compounded when that person is untrained or inadequately trained in performance assessment methodology. Finally, the immediate impact of assessment on the quality of instruction is greatly reduced when the feedback given to the performer is not communicated effectively and/or when the teacher is not given adequate opportunity to overcome weaknesses.

In fact, any one of these problems is enough to render an assessment of teacher performance indefensible in a technical and legal sense. But more importantly, poorly conducted performance assessments in the minimum competency evaluation context can lead to profoundly inappropriate impacts on teachers' lives and careers. Such high-stakes decisions demand only the best quality performance assessment.

Assessment for Professional Development. There are at least two instances when teacher performance might be assessed primarily for purposes of professional development. The first is at the time the teacher enters the profession. More and more, teacher induction is coming to mean a time when new staff members are

given the time and professional support needed to become accustomed to the school and classroom environment. Often this means the principal and/or senior staff will work with the new teacher to ensure the ongoing development of the teacher's instructional skills. Such support requires that the mentor rely heavily on observation and evaluation of classroom performance.

A second time when performance assessment for professional development becomes relevant is when the experienced and competent teacher decides to develop new professional capabilities—not because the district or a supervisor demands improvement, but because the teacher has high personal professional expectations. Under these circumstances, the teacher might well benefit from observation of and feedback on instruction, as he or she attempts to define new and important professional development goals. Further, that same teacher might continue to benefit from ongoing observation, evaluation and feedback, as she or he strives to improve. This kind of growth-oriented evaluation, although rarely the focus of teacher evaluation policy or publicity, represents one of the keys to the ongoing national efforts aimed at school improvement.

Performance assessment in the professional development context presents unique and interesting challenges. First, the objective is to provide information that will allow the teacher to move far beyond minimum competence toward excellence. For this reason, the performance criteria must reflect the teacher's personal commitment to improving. There is no requirement that criteria be standardized or public. Rather, the criteria can be individualized and very private if the teacher so wishes. In short, the criteria must relate directly to the professional development goal of the individual teacher.

Beyond the performance criteria, the assessment methodology available for use in this context is much more flexible than in the assessment of minimum competence. Observers may view naturally occurring events and documents, or the aspiring teacher may participate in courses and workshops that include structured exercises of various types. Depending on the situation, public or unobtrusive observations may be used and the amount of data gathered may range from a single observation to many observations over a period of time as long as several years. These will vary with the goal. Evaluation procedures are not constrained by law or contractual agreements. They are constrained only by the needs and desires of each individual teacher.

The rating of performance will need to be analytical so the developing teacher can track progress. The list of possible raters is long, including supervisors, peers, students (achievement and evaluations of teaching), and self-ratings. The list of possible recording strategies is similarly long, including checklists, rating scales, anecdotal records, and portfolios and tapes, again, depending on the teacher's professional development goal.

In this context, the keys to quality assessment are:

- selection of performance criteria that match the teacher's individual goal,
- the thoughtful observation of behavior and products that are similarly related to the teacher's goal,
- reliance on multiple trained observers, each of whom commands the respect of the teacher, and
- the careful delivery of enough detailed feedback to allow the teacher to adjust growth activities as needed to reach the goal.

Problems arise when the criteria fail to match the goal, the performance sampled fails to provide the needed information, observers lack credibility for the teacher, and/or feedback is delivered ineffectively.

When the purpose for performance assessment is the development of the professional competencies of an experienced and successful teacher, that teacher must be in charge of the evaluation. The teacher must lead the design of the assessment, identify the observers, review all data, and share results only if and when they wish to. But to take charge, teachers must be skilled in the use of performance assessment methodology on their own behalf.

Career Advancement. This decision context is like the others, in that it presents some unique challenges and requires very thoughtful use of assessment methodology. In this context, unlike the others, the task is to identify those who have attained the very highest levels of the performance continuum. They are tabbed to advance up the career ladder or to receive additional remuneration for outstanding performance. This decision often is made by a supervisor or a district management team.

The major challenge faced in this context is the development of appropriate performance criteria. Under some conditions, this may represent an insurmountable challenge. The potential prob-

lem is this: When we deal with minimum competencies, these can be defined and universally applied to all. That is, minimally acceptable performance for one teacher is generally the same as that of most teachers. But at the very high end of the performance continuum this may not be true. There can be many nitions of outstanding performance, depending on the performance context. Thus, there is no universally appropriate set of performance criteria. For instance, outstanding performance in an inner-city high school science classroom may differ fundamentally from outstanding performance in a suburban kindergarten. These differences may not simply be matters of degree. They may reflect differences in the kind of performance required to succeed.

This becomes a serious problem when we frame a decision context that requires that all teachers be measured against the same standard. Decisions as to who will receive merit pay and who will not require this, for example. Equality of opportunity demands that scarce merit-pay resources be made attainable to all on the same terms. But there may be few generalizable and job-relevant performance criteria upon which to base such a far-reaching competition. At the very least, the development of a universally applicable set of high-level performance criteria represents a formidable task that will demand considerable talent, time, and effort. One potential solution to this problem might be to group teachers into categories, such as grade levels, within which a common set of criteria can be applied and to award merit within categories only (i.e., allowing no cross-category comparisons based on differing, noncomparable criteria).

Other kinds of career-advancement decisions may not present such a difficult challenge. For instance, if a career ladder includes a position as a mentor teacher and a clear set of responsibilities is developed for the person who ascends to that level, then the performance criteria can be designed to assess attributes and skills relevant to that job (e.g., the ability to train adults). In any case, the primary key to success in conducting a valid and reliable performance assessment to identify outstanding performers lies in the difficult task of developing the right performance criteria.

In designing exercises for these assessments, either structured simulations or naturally-occurring events will serve. Most such assessments are based on the latter. Assessments often are public and rarely are unobtrusive. And they are most defensible when based on multiple observations of performance by qualified experts gathering data over a period of time. Qualified experts are

those who have been trained rigorously to apply carefully developed performance criteria.

ASSESSMENT PURPOSE AND THE KEYS TO SUCCESS

It is clear that different purposes require different kinds of assessment and different kinds of performance assessment. To bring this point home in very clear terms, let me summarize the foregoing discussion in a slightly different way. In the literature on teacher assessment, it is common for scholars and practitioners alike to draw the distinction between formative and summative uses of teacher-assessment results. Formative assessments, for example, serve to promote the professional development of teachers, whereas summative serve the district's personnel management needs. As a conclusion, let's contrast the keys to effective performance assessment in each of these general contexts.

Summative Assessments. Of the decision contexts previously discussed, several are summative in nature. These include final course and student-teaching assessments, certification, hiring, minimum competence assessment, and career advancement. In each of these cases, the first requirement is that the performance criteria be based on a thorough task analysis of the teaching process. This analysis will ensure the job relevance of the standards and will maximize the validity of the performance assessment. Further, sound assessment practice holds that teachers be told of the criteria by which they will be evaluated before the assessment takes place.

The second requirement is that the performance of the examinee be evaluated in the context of actual or simulated classroom settings. The sample of exercises—whether naturally occurring or structured exercises—must reflect in a representative manner the full range of situations in which the student or teacher will be expected to demonstrate proficiency when teaching. This too contributes to valid assessment.

The third requirement is that the raters of performance be thoroughly trained to apply the performance criteria in a systematic and consistent manner. This will minimize bias and increase the chances that a teacher's rating will reflect true capabilities rather than the idiosyncrasies of a particular judge. To further control for

bias it is advisable to have summative decisions be based on the observations and judgment of more than one judge whenever possible.

The fourth requirement of sound summative evaluation is that the teacher be given appropriate feedback on performance ratings and that the teacher have the opportunity to act upon that feedback to improve if necessary. This will afford the teacher the opportunity to complete professional development activities and repeat the performance assessment hopefully in a successful manner.

These four requirements are crucial because they protect the due process rights of teachers and provide the district with legally defensible evidence of proficiency for use in personnel actions. These are the standards by which we judge the quality of summative teacher assessments.

Formative Assessments. However, the standards by which we judge the quality of growth-oriented teacher assessments are quite different. Of the assessments just discussed, those that are of the formative variety are interim course and student teacher assessments, and professional development assessment at the time of induction or later in the competent teacher's career. In each of these cases, the test of the quality of the assessment is not its legal defensibility, but whether or not it contributes to the improvement of teacher performance. The keys to achieving this goal are fundamentally different than the keys to effective summative evaluation.

The first key is the teacher (for a more detailed discussion see Duke & Stiggins, 1986). To the extent that the teacher is knowledgeable about effective instruction, has high personal professional expectations, is open to criticism, is willing to change, and is comfortable with the material to be taught, the probability that a particular evaluation will result in growth for that teacher is greatly increased.

The second key to effective formative evaluation is the evaluator, or more specifically, how the teacher perceives the evaluator. The chances of a positive impact of evaluation increases when the teacher sees the evaluator as a credible source of ideas

---

*For a more detailed discussion of these, refer to Duke, D. L., & Stiggins, R. J. (1986). *Teacher evaluation: Five keys to growth.* Washington, DC: American Association of School Administrators, National Association of Elementary School Principals, National Association of Secondary School Principals, & National Education Association.

with persuasive rationale for those ideas. Furthermore, this evaluator must be perceived as patient and worthy of personal trust. Chances of a successful interchange are enhanced if the evaluator has a track record of helping people improve and is seen as sufficiently competent in his or her own right to take over the class and demonstrate the needed changes.

The third key to growth through assessment is the use of sound information gathering procedures. Performance criteria must be clear and perceived by the teacher as relevant in her or his specific classroom context. In addition, appropriate data sources need to be tapped including classroom observations, classroom document analysis, and student records. Finally, all relevant observers need to be tapped to ensure the reflection of a range of perspectives in classroom performance, including supervisors, colleagues, students, and self-assessments.

The fourth key to success in promoting teacher growth through assessment is the feedback provided to teachers. Feedback should be rich enough to act upon, but not so extensive that it overwhelms. It should be primarily descriptive, not only or harshly judgmental; sensitive to agreed upon professional development goals; formal and informal as the situation dictates; and, timed to promote effective communication (i.e., at a time when the teacher can listen and attend).

The fifth and final key to growth-oriented teacher assessment is that it be conducted in a district atmosphere in which teachers know that growth and improvement are valued and where resources (time and money) are appropriated to follow up the assessment with professional development activities and ongoing assessment.

Thus, the keys to successful performance assessment differ greatly as the overall purpose for assessment changes. This is precisely why one cannot carry out successful teacher assessment without a keen sense of purpose and a detailed knowledge of performance assessment methodology structure and design alternatives.

ACKNOWLEDGMENT

This chapter is based on work sponsored in part by the Office of Educational Research and Improvement (OERI), U.S. Department of Education under contract number 400-86-0006, with the Northwest Regional Educational Laboratory, Portland, OR. The content does not necessarily reflect the views of the department or any agency of the U.S. government.